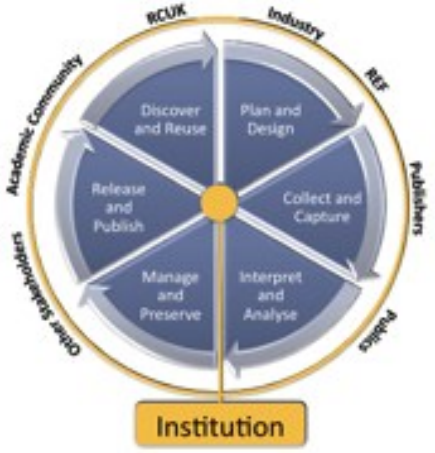
Managing Your Research Data

Catherine Pink (UKOLN)

Jez Cope (DTC)



The Research360 Institutional Research Lifecycle Concept

www.bath.ac.uk/rdso/datamanagement.html

After this workshop you will be able to:

- Understand what are research data and to whom do they belong
- Appreciate that the management and storing of research data are responsibilities of those who generate the information
- Learn about research data management strategies and tools
- Determine how much data needs to be managed
- Gauge for how long research data need to be maintained

What do you already know?

2 minute discussion topics – 1 per group

What do you understand these to mean?

Data

Metadata

Secure storage

Access to data

Intellectual property

A data management plan

What are Data?

- The lowest level of abstraction from which information and knowledge are derived

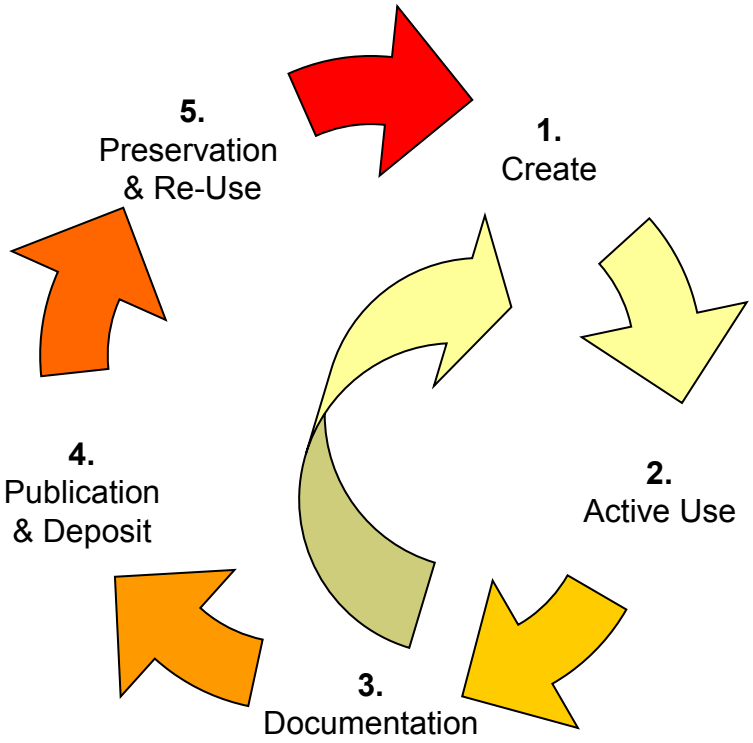
- Research data are collected, observed or created, for the purposes of analysis to produce and validate original research results

- Both analogue and digital materials are 'data'

- Digital data can be:
  - **created** in a digital form ("born digital")
  - **converted** to a digital form (digitised)


Write proposal
Start project
Acquire sample
Generate, Create, Collect
Process
Analyze
Interpret
Publish
Validate
Research Process

Data Lifecycles & Data Management Plans



1. What data will you produce?
2. How will you organise the data?
3. Can you/others understand the data
4. What data will be deposited and where?
5. Who will be interested in re-using the data?

Data Management Plans (DMPs)

DMPs are a framework:

- They ensure you've addressed all areas of data management
- DMPs do not check or validate your answers!

Required by the Code of Good Practice in Research

DMP Online (www.dmponline.dcc.ac.uk)

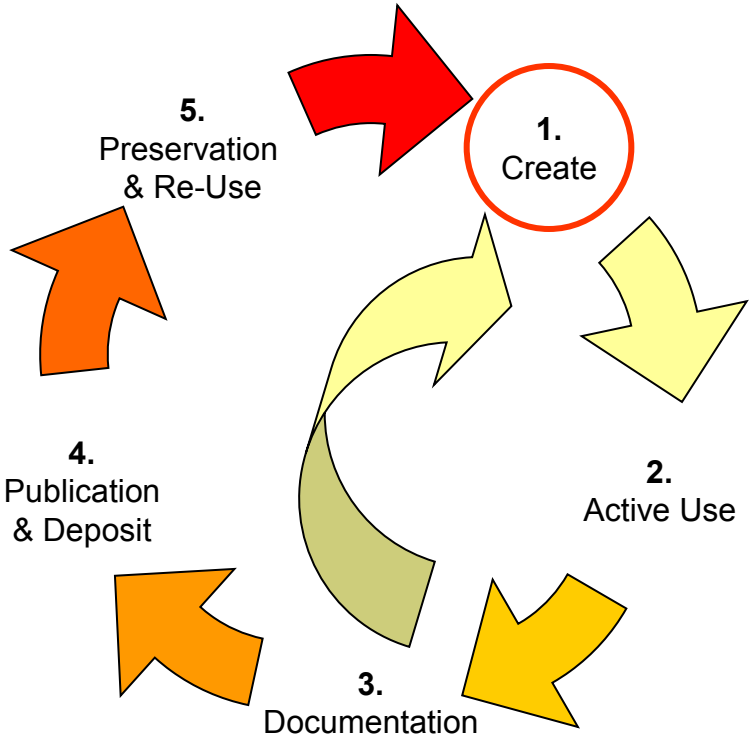
- Ideal for funding applications
- Based on the DCC checklist

DMPonline
The DCC Data Management Planning Tool

Postgraduate DMP template:

http://blogs.bath.ac.uk/research360/files//www/vhosts/bathblogs/wp-content/blogs.dir/969/files/2012/03/Data-Management-Plan-for-PGRs-v0.2.pdf

1. What data will you produce?



- What type of data will you produce?
- What types of file format?
- How easy is it to create or reproduce?
- Who owns it and is responsible for it?

Data Types

Data Type	Value	Example
Observational data captured around the time of the event	Usually irreplaceable	Sensor readings, telemetry, survey results, neuro-images
Experimental data from lab equipment	Often reproducible but can be expensive	Gene sequence, chromatograms, toroid magnetic field readings
Simulation data generated from test models	Model and metadata (inputs) more important than output data. Large modules can take a lot of computer time to reproduce	Climate models, economic models
Derived or compiled data	Reproducible (but very expensive)	Text and data mining, compiled databases, 3D models

Data can take many forms

- Notebooks & lab books
- Instrument measurements
- Experimental observations
- Still images, video & audio
- Survey results & interview transcripts
- Consent forms
- Text corpuses
- Models & software

Who owns or is responsible for your data?

Ownership

- Data ownership is complex, often defined on a case-by-case basis
- May be dependent on individual contractual agreements
- Contracts define needs of the University, staff, students, funders, collaborators

Management

- Defined in the University of Bath **Code of Good Practice in Research**: http://www.bath.ac.uk/opp/research/

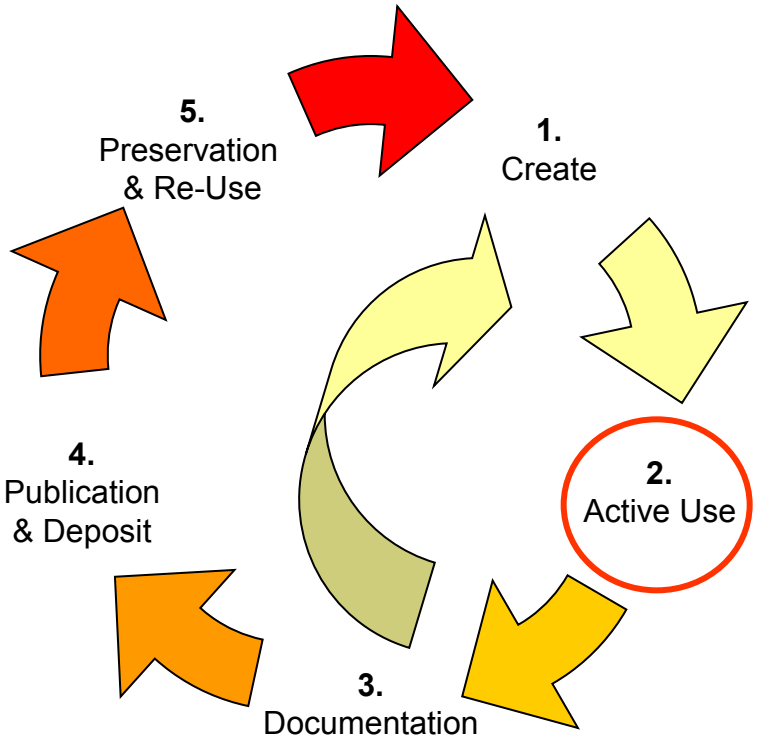
Who owns or is responsible for your data?

In practice

- Everyone plays their part
- If you're generating and using data, you should:
  - Comply with guidelines from your group, department, faculty, collaborators
  - Make sure your data is securely stored and backed up
  - Describe your data so that you/others can understand it in future
- If you're managing a project, you should:
  - Be fully aware of funder, collaborator and publisher requirements
  - Ensure you have access to group data
  - Assess what should be published and/or archived

More info: http://www.data-archive.ac.uk/create-manage

2. How will you look after your data?



- Is your data safe?
- Is your data organised?
- Can you find your data?

Storage and Security

Photo credits: Harvey Rutt
http://www.ecs.soton.ac.uk/regenesis/pictures/

3… 2… 1… Backup!

- at least 3 copies of a file
- on at least 2 different media
- with at least 1 offsite

Test file recovery

- At set up time and on a regular basis

Access

- Protect your hardware
- If sensitive use file encryption
- Keep passwords safe (e.g. Keypass)
- At least 2 people should have access to your data

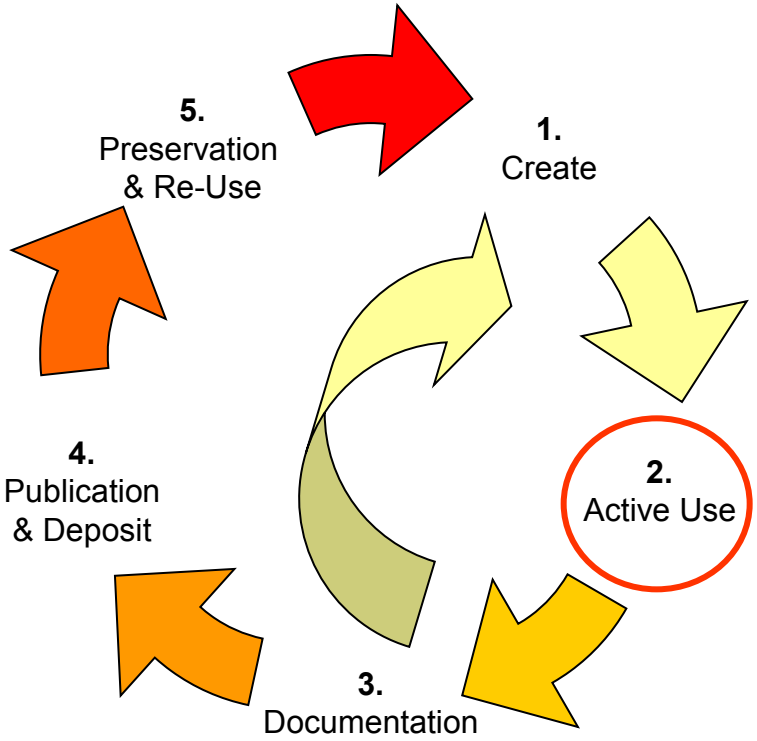
Ask BUCS for advice

More info: http://www.data-archive.ac.uk/create-manage/storage

Storage and Security – Back up options

Media	Advantages	Disadvantages
CDs or DVDs	• Useful for quick restore in the event of minor disaster	• Static capture of data • Not built to last • Vulnerable to theft • Physical loss of media
External hard drives	• Dynamic capture of data • Useful for quick restore in the event of minor disaster	• Must store securely and remotely to original copy • Vulnerable to theft • Must use file encryption if sensitive
BUCS server	• Resilient backup • X:drive (1Tb free per project) • Safety net for major disaster	• Lack of offline access • Must have a BUCS account
Digital scans of lab books	• Easy to do on a daily basis at any campus printer • Automatically save daily page scans to your H:drive	• Manipulation of page content difficult

2. How will you look after my data?



- Is your data safe?
- Is your data organised?
- Can you find your data?

Can you find your data?
If not, have you considered…

A Clear Directory Structure

- Top level folder and substructure

File Version Control

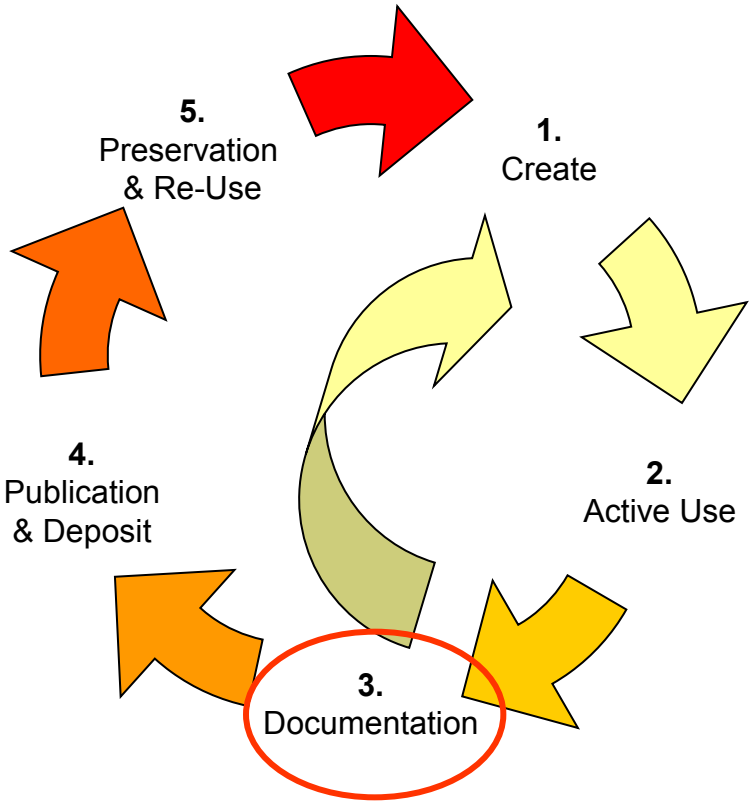
- Discard obsolete versions if no longer needed after making backups
- Manage using: File naming (see below)
  Version control software (e.g. Git, Mercurial, SVN)

File Naming Conventions

http://www.jiscdigitalmedia.ac.uk/crossmedia/advice/choosing-a-file-name/

- Record any naming conventions or abbreviations used
  - E.g. [Experiment]_[Reagent]_[Instrument]_[YYYYMMDD].dat
- Date/time stamp or use a separate ID (e.g. v1) for each version

3. Documenting data



- Do you still understand your older work?
- Is the file structure / naming understandable to others?
- Which data will be kept?
- Which data can be discarded?

Understanding your data

- Students:
  - Will you be able to write up your methods at the end of your studies?
- Project leads:
  - Will you be able to respond to reviewers comments?
  - Will you be able to find the information you need for final project reports?
- Can you reproduce your work if you need to?
- What information would someone else need to replicate your work?

Understanding your data

Do you know how you generated your data?

- Equipment or software used
- Experimental protocol
- Other things included in (e.g.) a lab notebook
- Can reference a published article, if it covers everything

Are you able to give credit to external sources of data?

- Include details of where the data are held, identified & accessed
- Cite a publication describing the data
- Cite the data itself e.g.

10.5255/UKDA-SN-7018-1

Citation:
ScotCen Social Research, *Scottish Social Attitudes Survey, 2010* [computer file]. Colchester, Essex: UK Data Archive [distributor], May 2012. SN: 7018, http://dx.doi.org/10.5255/UKDA-SN-7018-1

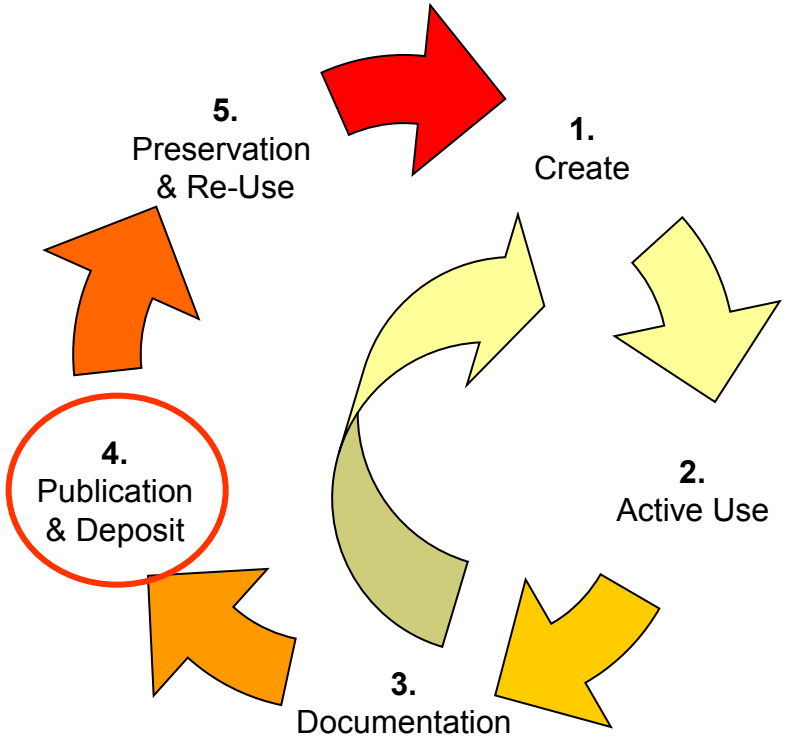
Change log:
11 May 2012: DOI created

Metadata

- Contextual information for data is called metadata — literally data about data
- Data repositories & archives require some generic metadata, e.g.
  - author, title, publication date
- For data to be useful, it will also need subject-specific metadata e.g.
  - reagent names, experimental conditions, population demographic
- Record contextual information in a text file (such as a 'read me' file) in the same directory as the data e.g.
  - codes for categorical survey responses
  - '999 indicates a dummy value in the data'

More info: http://www.data-archive.ac.uk/create-manage/document

4. What data will be deposited and where?



- Are you *expected* to share your data?
- Are you *allowed* to share your data?
- Define the core data set of the project
- Which data will be included in your publication / thesis?

Data Sharing – Why share your data?

- Share with your future self – avoid repeating research!
- Promote your research – get cited!
- Enable new discoveries
- Replication
- Store your data in a reliable archive
- Comply with funding requirements

Requirements to share your data

Some journal publishers have a policy on data availability:

- Are you making any of your data available as supplementary information?
- Is there sufficient information with the data so that it can be understood and reused?

Most UK funders now expect research data to be made publically available

RESEARCH COUNCILS UK

Common Principles on Data Policy

"Publicly funded research data are a public good, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner that does not harm intellectual property."

Find your funder's policy: http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies

Restrictions on sharing your data

Are there privacy requirements from the funders or commercial partners?

- e.g. personal data, high security data

You might not have the right to share data collected from other sources

- It depends upon whether those data were licensed and have terms of use
- Most databases are licensed and prohibit redistribution of data without permission

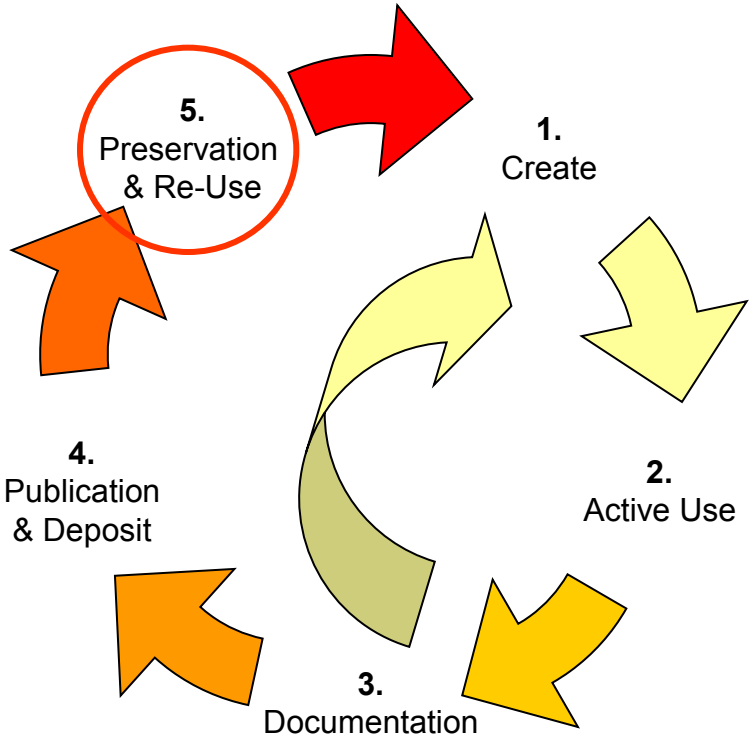
If you are uncertain as to your rights to disseminate data, contact research-data@bath.ac.uk

How to share your data

- Deposit in a data repository eg. GenBank
- Data can be licensed
- Culture of data sharing: can make available your data under a CC-BY or CC0 declaration to make this explicit
    - CC-BY license permits reuse but requires attribution
    - CC0 declaration is a waiver of copyright.
    - Laws about data vary in different countries.
- You may have rights to first use or to commercial exploit data
- How to license research data:

  http://www.dcc.ac.uk/resources/how-guides/license-research-data

5. Preservation and Re-use



- How long will your data be reusable for?
- Do you need to prepare your data for long term archive?
- Which data do you need to keep?

Data retention and archiving

How permanent are the data?

- Short term (e.g. 3-5 years)
- Long term (e.g. 10 years)
- Indefinite

Should discarded data be destroyed?

- Keep all versions? Just final version? First and last?

What are the re-processing costs?

- Keep only software and protocol/methodology information

Are there tools/software needed to create, process or visualise the data?

- Archive these with your data

File formats for long-term access

- Unencrypted
- Uncompressed
- Non-proprietary/patent-encumbered
- Open, documented standard
- Standard representation (ASCII, Unicode)

Type	Recommended	Avoid for data sharing
Tabular data	CSV, TSV, SPSS portable	Excel
Text	Plain text, HTML, RTF PDF/A only if layout matters	Word
Media	Container: MP4, Ogg Codec: Theora, Dirac, FLAC	Quicktime H264
Images	TIFF, JPEG2000, PNG	GIF, JPG
Structured data	XML, RDF	RDBMS

Further examples: http://www.data-archive.ac.uk/create-manage/format/formats-table

Summary

- Data management is important at all stages of a project
- There are tools available to help you
- Keep your data safe
  - Back up your data
  - Test your back-ups
- Keep your data organised
  - Find it – good formats and file names
  - Understand it - check documentation and metadata
- Consider publishing your data so that you can get recognition for your work
- Ask for help: research-data@bath.ac.uk

What do you think now?

Further Information

- http://go.bath.ac.uk/research-data
- research-data@bath.ac.uk

Contacts

Cathy Pink
Institutional Data Scientist
c.j.pink@bath.ac.uk

Jez Cope
Technical Data Coordinator
j.cope@bath.ac.uk

Acknowledgements

This course was based on:

Research Data MANTRA [online course]
Created by EDINA and Data Library, University of Edinburgh
Available at http://datalib.edina.ac.uk/mantra

DataTrainArchaeology: Teaching Material Downloads

and work done by:

- Professor Richard H. Guy, Dept. of Pharmacy & Pharmacology
- http://libraries.mit.edu/data-management